**Exercise Z**

*Using Apache Zeppelin to manage your Spark Experience*

**Prior Knowledge**

Previous Apache Spark lessons

**Learning Objectives**

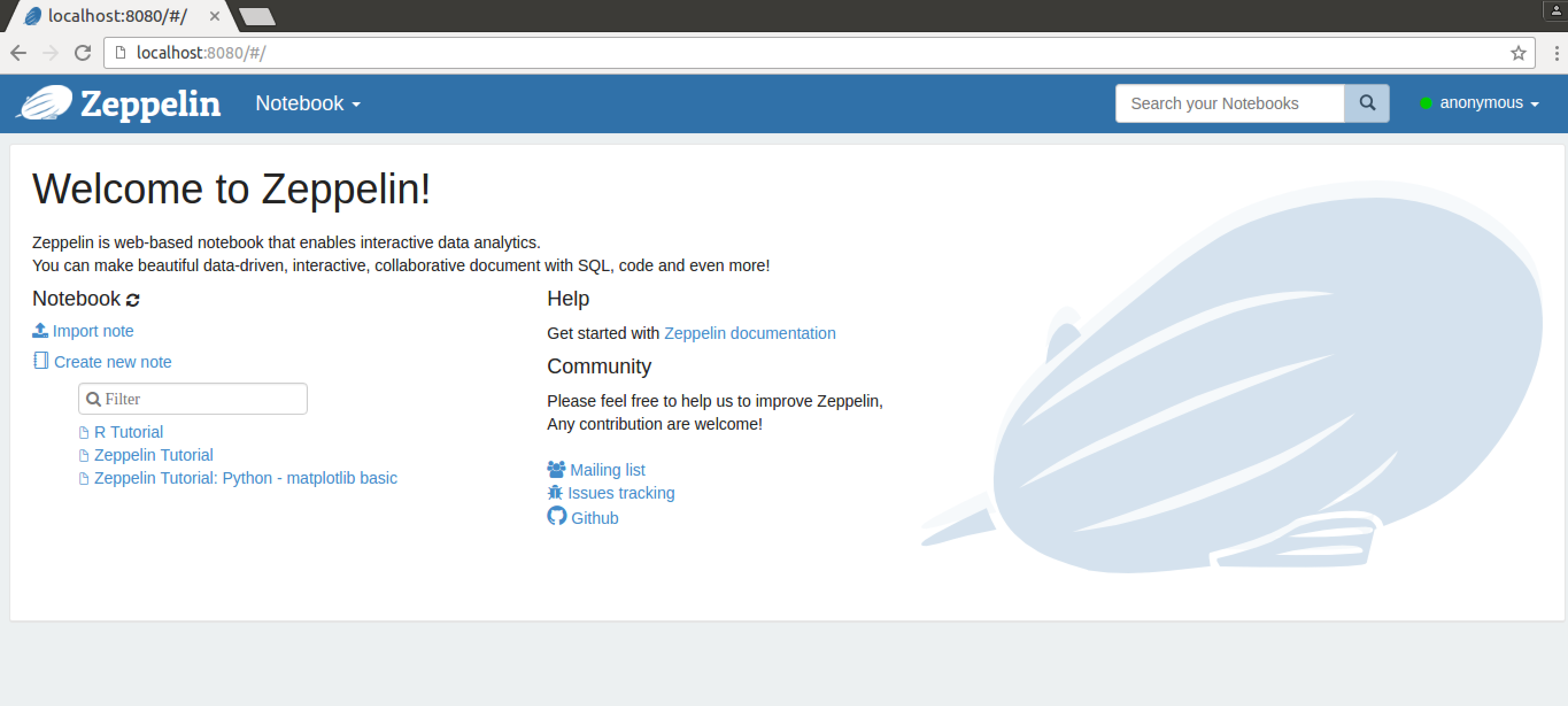
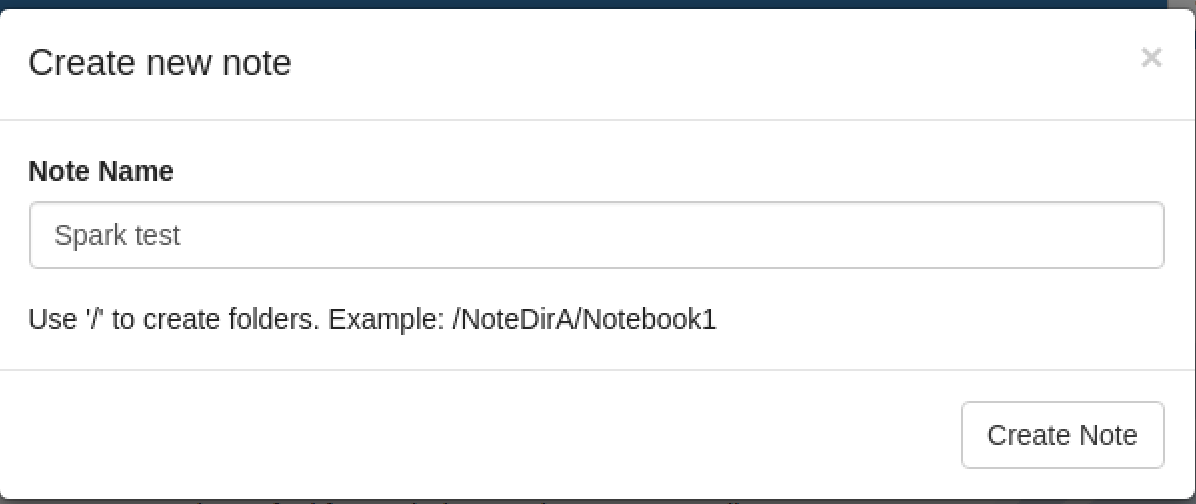
Using a notebook

**Software Requirements**

(see separate document for installation of these)

* Apache Spark 2.0.0
* Python 2.7.x
* Nano text editor or other text editor

**Steps**

1. Install Zeppelin:  
   curl -L <http://freo.me/zeppinst> | sh  
   (Wait a bit - it’s a fairly large download)
2. Tell Zeppelin to use CSV package: (All on one line)  
   export SPARK\_SUBMIT\_OPTIONS="--packages com.databricks:spark-csv\_2.10:1.2.0"
3. Start Zeppelin  
   cd ~/zepp  
   bin/zeppelin-daemon.sh start  
     
   You should see:  
   Log dir doesn't exist, create /home/oxclo/zepp/logs  
   Pid dir doesn't exist, create /home/oxclo/zepp/run  
   Zeppelin start [ OK ]
4. Browse to <http://localhost:8080>  
   You should see:  
   
5. Click on **Create New Note**
6. Give it a name:
7. From this page <http://freo.me/zepp-pc> paste the following:

%pyspark

sqlc = SQLContext(sc)

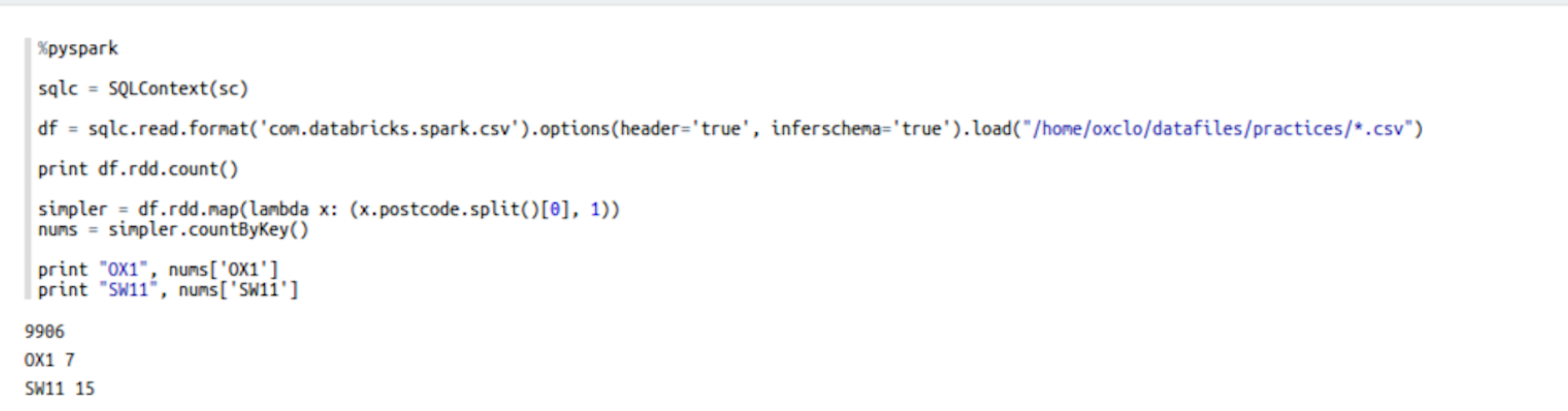
df = sqlc.read.format('com.databricks.spark.csv').options(header='true', inferschema='true').load("/home/oxclo/datafiles/practices/\*.csv")

simpler = df.rdd.map(lambda x: (x.postcode.split()[0], 1))

nums = simpler.countByKey()

print "OX1", nums['OX1']

print "SW11", nums['SW11']

1. The first line is a hint to Zeppelin that it needs to use the Pyspark interpreter (Zeppelin supports multiple different backends)
2. Hit the Run button (the little “Play Arrow”)
3. You should see:  
   
4. That’s all!